ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Exploring the Application of Artificial Intelligence in Tax Risk Identification and Management in the Context of Big Data

Xiaowen Shen[1,*]

[1] Economics and Management School, Sichuan Polytechnic University, Deyang, Sichuan, 618000, China

SUMMARY: *How to effectively respond to tax risk identification and management challenges amid the proliferation of big data is an urgent problem for enterprises. In this study, the data sources are firstly clarified, and the redundancy and noise are eliminated through preprocessing, which ensures the data quality and enhances the credibility of the subsequent conclusions. The features of enterprise tax risk management are imported into SVM as inputs for identification and prediction, and the traditional SVM algorithm takes a long time to compute when facing a large amount of feature data. To address this issue, the least squares (LS) method is introduced to optimize the conventional SVM framework, converting the quadratic programming challenge into a system of linear equations, thereby maximizing the identification performance, and finally obtain a LS-SVM-based tax risk management identification algorithm. -SVM-based recognition model for tax risk management. With the theoretical support of research dataset, confusion matrix, and quantitative assessment indexes, the model is deeply explored and analyzed. The LS-SVM hybrid algorithm achieves an accuracy of 0.9772, precision of 0.9778, recall of 0.9911, F1 score of 0.9844, and AUC of 0.9922, all of which surpass the performance of the decision tree, random forest, and logistic regression approaches in recognition, interpreting the application value of LS-SVM algorithm in tax risk identification and management.*

KEYWORDS: *LS-SVM; confusion matrix; data preprocessing; tax risk management; identification model*

1 Introduction

In the rapidly evolving landscape of the digital economy, enterprises face a variety of tax risks, firstly, whether the enterprise's tax behavior is legal and compliant, whether there is a situation in which the enterprise should pay tax but not pay tax or pay less tax, which leads to the risk of back taxes, fines, and loss of enterprise credit or reputation [1, 2]; secondly, whether the enterprise fully utilizes the relevant policies of the tax incentives, and reasonably carries out tax planning according to the law, so as to prevent the assumption of unnecessary taxes [3]; Thirdly, whether the enterprise intentionally or unintentionally fails to follow the requirements of tax laws and regulations for tax declaration, invoice management and so on. The traditional enterprise tax risk identification and management intelligence relies on the enterprise to combine its own actual business to start, and the development of artificial intelligence (AI) provides support for the intelligence of enterprise tax risk identification and management.

AI can accurately identify the tax risk signals of enterprises through the collection and

*aw86878889aw@163.com
https://doi.org/10.65102/is2026143
Published: 30 April, 2026

analysis of massive data [4]. Through the establishment of data models and risk early warning mechanisms, AI can monitor the operating conditions and financial situation of enterprises in real time, and discover potential risk points and behavioral anomalies [5]. At the same time, AI can also identify possible false transactions and tax evasion through advanced techniques such as pattern analysis and knowledge extraction to improve the effectiveness and accuracy of tax management [6, 7]. In addition, AI demonstrates the capacity to evaluate corporate tax exposure by leveraging large-scale datasets and intelligent algorithms [8]. By carefully analyzing and comparing tax data, AI can assess an enterprise's ability to pay taxes, its tax liability, and its tax compliance [9, 10]. AI can also predict and control the tax risks that may exist in the future of an enterprise through simulation prediction and risk assessment models [11].

Regarding the research on how AI is applied to tax risk identification and management and its impact, literature [12] analyzed the application of AI in tax risk management, pointed out that it can assist in the identification of high-risk cases in order to prevent tax evasion, and emphasized that improving the legal supervision and control mechanisms of AI systems is a key challenge to ensure their effective and legal operation. Literature [13] discusses the positive role of AI in tax risk identification and management, and by analyzing its ability to process complex data and identify patterns, it points out that AI can significantly improve the efficiency of taxpayer classification and fraud detection, and emphasizes its important value in optimizing debt collection, assisting decision-making, and building a fair and transparent tax environment. Literature [14] proposes a deep learning method named FLO-IG-LSTM to enhance tax risk identification, analyzes the significant advantages of the model in improving the detection accuracy and operational efficiency by integrating multi-source historical data and utilizing linear discriminant analysis for feature extraction, and emphasizes its positive role in promoting the automation and intelligence of tax risk management. Literature [15] explored the combination of AI technology and tax risk management, and analyzed its positive role in improving the accuracy of risk prediction, reducing the tax cost and enhancing the reliability of data by integrating algorithms such as bag and support vector machine. Literature [16] investigates the role of AI in tax risk identification and management, pointing out its ability to optimize tax burdens, enhance anomaly detection precision, and shorten the time to compliance through machine learning and natural language processing, while also exploring its potential to automate transfer pricing assessment and assist in managerial decision processes, and emphasizing the importance of establishing an ethical framework and data governance to take advantage of its benefits. Literature [17] analyzes the application of AI in tax risk identification and management, noting its ability to effectively address tax gap challenges and protect lost revenues during audits, and emphasizes the positive role of the technology in improving tax administration efficiency and maintaining tax fairness. Literature [18] takes Indonesia as an example, revealing that AI can assist law enforcement, enhance tax convenience and fairness, and reduce compliance costs, while examining the country's advantages in terms of technological openness and strategic support, and pointing out barriers to its application such as the lack of regulations and insufficient human resources.

In addition, literature [19] explored the actual effect of AI in tax control to reduce information risk through correlation analysis, pointing out that it is statistically significant in improving risk management and cybersecurity, and can enhance the reliability of the data, and analyzed the direction of improvement in the information verification and other aspects of the information needs to be further optimized in order to achieve the depth of integration. Literature [20] emphasized that the effective use of digital tools, emerging communication channels and data resources has significantly improved the efficiency of tax services in recent years.

Literature [21] proposed an intelligent decision-making system based on neural network and cognitive modeling for identifying corporate tax evasion risks, analyzed its effectiveness in enhancing tax transparency and meeting the challenges of digital supply chains by constructing an information framework of potentially suspicious events, and verified the practical value of the approach by taking the example of supply chains in the Arctic region of Russia. Literature [22] reviewed the pivotal function of AI in ensuring tax compliance and strengthening risk detection, and by examining the capacity of machine learning algorithms to handle vast volumes of financial data, pointed out that they can efficiently identify fraud patterns, predict taxpayer behaviors, and optimize law enforcement resources, thus improving the efficiency of tax administration and reducing the burden of compliance, and at the same time, emphasized the need to build a robust governance framework to ensure its transparent and accountable application. Literature [23] analyzes the comprehensive impact of AI in the tax field, exploring its contribution to tax compliance, fraud identification and service optimization through descriptive data, and examining its potential to increase tax revenue through specific income tax provisions, while pointing out the need to pay attention to the potential negative impact on the income of some professions, thus providing a reference for balanced development of multi-sectoral policymaking. Literature [24] applied an interpretable AI model to predict the level of tax administration in the manufacturing industry, pointed out that the Random Forest model was optimal in terms of accuracy and generalization ability by comparing multiple algorithms, and enhanced the interpretability of the results by combining Shapley's addition and interpretation techniques, analyzed the impact of key characteristics on the tax burden ratio, and thus provided data support for managers to formulate sustainable tax decisions.

Feature screening, outlier screening, missing value processing, balanced data set, data set slicing, and data standardization are carried out on the raw data before mathematical modeling, so that the results of its research have stronger explanatory power. On the basis of the theory of artificial intelligence technology, the SVM algorithm is selected to carry out the enterprise tax risk management identification process, there is a problem of long computation time, in this regard, the use of least squares (LS) to improve the traditional SVM algorithm, so as to solve the quadratic programming problem into solving a set of linear equations, to maximize the reduction of algorithmic computation time, which designed a LS-SVM-based tax risk management identification model. Finally, the data preprocessing and model validation analysis are carried out under the joint effect of data set, evaluation indexes and confusion matrix, aiming to reveal the application value of artificial intelligence in tax risk identification and management under the background of big data.

2 Exploring Tax Risk Identification and Management

2.1 Data sources

The raw data used in this study are obtained from the Cathay Pacific database, and the relevant financial statement data and tax violations of Chinese listed manufacturing enterprises are selected to summarize and constitute the initial analytical dataset from 2018-2023. This study summarizes and analyzes the collected data through Rstudio software, and calculates new characteristic variables based on the derivation of some characteristic variables, which are used to reflect the financial status of enterprises.

2.2 Data pre-processing

Data preprocessing encompasses the procedures of purifying, reshaping, and consolidating the original data before data analysis, and the main purpose of data preprocessing is to make the data more reliable and effective in the subsequent analysis or model construction.

(1) Characteristic variable screening

After determining the initial dataset source for this research, the sample raw dataset contains 12 characteristic variables, detailed specific names in order: cash ratio (N1), tangible assets ratio (N2), fixed assets ratio (N3), current assets ratio (N4), non-current assets ratio (N5), current assets turnover ratio (N6) , earnings per share (N7), diluted earnings per share (N8), net operating cash receipts (N9), equity ratio (N10), operating cycle (N11), and management expense ratio (N12), and named them as N1, N2, N3, N4, N5, N6, N7, N8, N9, N10, N11, and N12.Since the original data As more feature variables are composed of other feature variables in ratio of two by two, putting them together in the model may lead to the problem of multicollinearity of feature variables, the Spearman correlation coefficient is employed to quantify the degree of association among the feature variables. By computing the pairwise correlation values and removing variables exhibiting strong interdependence (Spearmanr>0.8) according to their size, it not only makes the interpretation of the characteristic variables stronger, but also improves the performance of the model. Therefore, considering the accuracy, effectiveness and simplicity of the model constructed by the later study at runtime, this paper screens 12 feature variables.

(2) Outlier screening

The data utilized in this study originates from publicly listed firms within the manufacturing sector, which encompasses numerous sub-industries, and notable discrepancies exist across different sample enterprises. Moreover, the characteristic variables in the dataset of this study are more synthetic variables, i.e., based on the processing of the original financial data, and the distribution is more discrete and does not obey normal distribution, so it is not possible to screen the outliers through the box-and-line diagram or the distribution situation. Therefore, in order to ensure the completeness and authenticity of the indicators of each characteristic variable, considering the impact of zero value on the bias of the final prediction model, the zero value is deleted as an outlier.

(3) Missing value processing

Most of the feature variables in the initial dataset have missing values, and the processing of missing values can start from the sample or from the feature variables. When a feature variable corresponds to a large number of missing values, it is easy to cause data distortion by filling a large area of it, so this kind of situation is usually the eradication of the feature variable, and when a feature variable corresponds to a small number of missing values, it can be filled by interpolation, quadratic filling method and so on to fill the missing values or directly delete the samples with missing values.

(4) Balanced data set

Unbalanced dataset mainly refers to the large difference in the number of each category of the response variable, and the serious imbalance problem may cause the machine learning algorithm to learn more about the samples with a larger proportion of the category when training the model, thus ignoring the situation of a few categories. After SMOTE processing, the proportion of majority-class samples in the dataset reaches approximately double the number of minority samples, effectively mitigating the potential impact of the data imbalance problem on subsequent model training.

(5) Data set slicing

Furthermore, separating the training set and test set can effectively avoid the overfitting problem, which means that the model performs well on the training set, but performs poorly on

the unknown data. By dividing the dataset into training set and test set, the training parameters of the model on the training set can be adjusted at any time during the training process to prevent overfitting. Samples are allocated using stratified random sampling based on class labels, yielding a training-to-test partition of 7:3.

(6) Data Standardization

In the dataset used in this study, all feature variables are of ratio type and the ratios are converted to decimal form. Due to the existence of extreme values within the data, such data are too large or too small, which will increase or weaken the bias and weight of the feature variables under the sample, thereby compromising the predictive accuracy of the resulting model, so it is necessary to standardize the data set. Currently, the commonly used data standardization methods are polar deviation standardization method and Z-Score standardization method. Considering that the feature variables in the dataset do not obey the normal distribution, and there are more outliers, the calculation using the extreme difference method can not effectively eliminate the influence of the magnitude, so this paper adopts the Z-Score method for standardization processing. Its standardization processing formula is:

$$N' = \left(N - \bar{N}\right) / \sigma \tag{1}$$

where N is the sample input value, $\bar{N}$ is the sample mean and σ is the standard deviation, so the method is also known as standard deviation normalization method. In the R language, the standardization of the data set will be achieved through the scale function. Since the response variable is only used as an identifier for classification, it will be excluded in the data standardization. The variation range of the processed feature variable data is $[-1,1]$.

2.3 Recognition model for tax risk management based on LS-SVM

The previous research work explores tax risk identification and management research data sources and pre-processing to ensure the availability of research data, while also identifying nine features of tax risk management. The features are imported into SVM as inputs for identification and prediction, and the traditional SVM algorithm leads to a longer computation time in the face of more feature data, for this reason, it is proposed to improve the traditional SVM algorithm by applying the method of least squares (LS), so as to transform the problem of solving quadratic programming into solving a set of linear equations, maximize the identification performance, and finally obtain a recognition model based on LS- SVM-based tax risk management recognition model.

2.3.1 SVM Theoretical Foundations

Assume that there exists an unknown functional relationship linking the output variable y and the input variable x,which can be theoretically characterized as a joint probability distribution $F(x, y)$ defined over the two variable spaces. Machine learning is the use of n independent and identically distributed samples of observations:

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n) \tag{2}$$

The objective is to determine an optimal mapping $f(x, w)$ serving as an approximation of the underlying dependence within a family of functions $\{f(x, w)\}$ that minimizes the expected prediction risk $R(w)$. That is:

$$Minimize \quad R(w) = \int L(y, f(x,w)) dF(x,y) \tag{3}$$

where $\{f(x,w)\}$ is the set of prediction functions, which can be expressed as any function; w is the generalized parameter of the function; and $L(y, f(x,w))$ is the loss due to the prediction of y by $f(x,w)$. Given that the explicit form of the joint probability distribution $F(x,y)$ remains unavailable, direct computation of the expected risk minimizer through the above equation is infeasible, and one viable strategy is to approximate it from observed samples by invoking the law of large numbers. Thus equation (4) can be defined to approximate the expected risk. That is:

$$R_{emp}(w) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i, w)) \tag{4}$$

Since $R_{emp}(w)$ is defined using known training samples, it is called empirical risk. The so-called empirical risk minimization (ERM) principle is to replace the minimum expected risk with the minimum empirical risk estimated from the sample, and the neural network model embodies this idea.

According to statistical learning theory, the following bound relating the empirical risk $R_{emp}(w)$ and the true risk $R(w)$ holds with probability no less than $1-\eta$:

$$\begin{aligned} R(w) &\le R_{emp}(w) + \sqrt{\frac{h(\ln(2n/h+1) - \ln(\eta/4))}{n}} \\ &= R_{emp}(w) + \phi(h/n) \end{aligned} \tag{5}$$

where h is the VC dimension of the function set, the VC dimension of the m-dimensional space is $m+1$, and n is the number of samples.

Assume that the training samples $\{x_i, y_i\}$, $i = 1, \cdots, n$, $x_i \in R^n$, $y_i \in \{-1, +1\}$, if there exists a categorical hyperplane $w \cdot x + b = 0$ such that:

$$\begin{aligned} &w \cdot x + b \ge 1, \ y_i = 1 \\ &w \cdot x + b \le -1, \ y_i = -1, i = 1, \cdots, n \quad w \in R^n \end{aligned} \tag{6}$$

Under these conditions, the training set is said to be linearly separable, where $w \cdot x$ represents the inner product of vectors $w \in R^n$ and $x \in R^n$. Equation (6) can be written in the following form:

$$y_i \left(w \cdot x_i + b \right) \ge 1, \quad i = 1, 2, \cdots, n \quad w \in R^n \tag{7}$$

For the linearly differentiable case, the point x_i in each class is at a distance from the hyperplane $w \cdot x + b = 0$:

$$d_i = \frac{w \cdot x_i + b}{\|w\|} \tag{8}$$

can be obtained from Eqs. (7) and (8):

$$y_i d_i \geq \frac{1}{\|w\|} \tag{9}$$

Thus, $\frac{1}{\|w\|}$ is the lower bound on the shortest distance between the point x_i in each class and the hyperplane. Therefore, in the linearly differentiable case, the solution of a support vector machine can be transformed into solving the following linear programming problem:

$$\begin{aligned} &Minimize \ \frac{1}{2}\|w\|^2 \\ &S.t. \ y_i\left(w \cdot x_i + b\right) \geq 1, \quad i = 1, 2, \cdots, n \end{aligned} \tag{10}$$

After finding the optimal solution (w, b), the corresponding optimal hyperplane equation is:

$$w \cdot x + b = 0 \tag{11}$$

This can be categorized by the following decision function:

$$f(x) = \text{sgn}(w \cdot x) \tag{12}$$

where $\text{sgn}(\ )$ is the sign function.

In practice, most of the cases are nonlinearly differentiable, and since it is not possible to construct a categorical hyperplane as in the linearly differentiable case, it is necessary to introduce a nonnegative slack variable ξ_i and the corresponding penalty coefficients C to transform the model, and in the linearly indivisible case, the solution of the support vector machine is transformed into solving the following linear programming problem:

$$\begin{aligned} &Minimize \ \frac{1}{2}\|w\|^2 + C\sum \xi_i \\ &S.t. \ y_i\left(w \cdot x_i + b\right) \geq 1 - \xi_i, i = 1, 2, \cdots, n \ \ \xi_i \geq 0 \end{aligned} \tag{13}$$

where ξ_i is the distance of the sample point (x_i, y_i) from the class, which can be regarded as the deviation of the sample point about the classification hyperplane, and C is the penalty coefficient, and the larger C indicates the larger the penalty for misclassification. Generally speaking, the model solution process will be through the original problem into its dual problem, using the original objective function and constraints to establish the Lagrange function for the operation, the original problem of the Lagrange function is:

$$\begin{aligned} L = &\frac{1}{2}\|w\|^2 + C\sum_{i=1} \xi_i - \sum_{i=1} \alpha_i \left[y_i\left(w \cdot x_i + b\right) - 1 + \xi_i \right] \\ &- \sum_{i=1} \beta_i \xi_i \end{aligned} \tag{14}$$

Make pairwise transformations to it:

$$\begin{aligned} &Maximize\ \sum_{i=1}\alpha_i-\frac{1}{2}\sum_{i,j=1}\alpha_i\alpha_j x_i x_j y_i y_j \\ &S.t.\ \sum_{i=1} y_i\alpha_i=0\,;0\le\alpha_i\le C \end{aligned} \tag{15}$$

From this, we can find the optimal w and b, viz:

$$w=\sum_{i=1}^{n}\alpha_i y_i x_i \tag{16}$$

$$b=y_i-w\cdot x_i \tag{17}$$

It's available:

$$\alpha_i\left[y_i\left(wx_i+b\right)-1+\xi_i\right]=0 \tag{18}$$

$$\left(C-\alpha_i\right)\xi_i=0 \tag{19}$$

For α_i it can be discussed in 3 cases:

(1) $0<\alpha_i<C$. From Eq. (19), we can get $\xi_i=0$; from Eq. (18), we can get $y_i\left(wx_i+b\right)-1=0$. At this point, x_i is the support vector.

(2) $\alpha_i=C$. At this point, from equation (19), we can get that $\xi_i=0$ or $\xi_i\neq 0$. When $\xi_i\neq 0$, it is obtained from Eq. (18) that $y_i\left(wx_i+b\right)<1$, at this time the corresponding point is a misjudgment point. When $\xi_i=0$, the corresponding point is the support vector.

(3) $\alpha_i=0$. At this point, $\xi_i=0$ can be obtained from Eq. (19), and Eq. (18) holds constant. In this case, actually the constraints do not work, i.e., the corresponding points are correctly classified.

Assuming that the input sample x_i is mapped by the mapping function $\phi(\cdot)$ into the high-dimensional space $\phi(x_i)$, and the problem is solved in this high-dimensional space using the linear classification function, according to the principle of minimization of structural risk, the optimal hyperplane equation $w^T\phi(x)+b=0$ can be found out through the solution (w,b) of the following model:

$$\begin{aligned} &Minimize\ \frac{1}{2}\|w\|^2+C\sum_{i=1}\xi_i \\ &S.t.\ y_i\left\lfloor w^T\phi(x_i)+b\right\rfloor\ge 1-\xi_i,\ \xi_i\ge 0 \end{aligned} \tag{20}$$

In that case, the derivation of the problem is transformed into a linear classification problem, except that x_i is replaced by $\phi(x_i)$, and the dyadic problem in high-dimensional space is as follows:

$$\begin{aligned} &Maximize\ \sum_{i=1}\alpha_i-\frac{1}{2}\sum_{i,j=1}y_iy_j\phi(x_i)\cdot\phi(x_j)\alpha_i\alpha_j \\ &S.t.\ \sum_{i=1}\alpha_iy_i=0,0\le\alpha_i\le C \end{aligned} \tag{21}$$

Without knowing the exact form of $\phi(x_i)$, it is only necessary to pass the kernel function K in the original space such that $K(x_i,x_j)=\phi(x_i)\cdot\phi(x_j)$, and to avoid the feature space dimension by having the inner-product computation accomplished by the kernel function in sample space, thus avoiding the feature space dimension catastrophe problem.

The kernel function can be regarded as a definition of the distance between samples, and different kernel functions are used to produce different support vector machines. The main types of kernel functions commonly used are as follows:

(1) Linear kernel function: $K(x_i,x_j)=(x_i,x_j)$.

(2) Polynomial kernel function: $K(x_i,x_j)=\left[(x_i,x_j)+1\right]^d$, d is a natural number.

(3) Radial basis kernel function: $K(x_i,x_j)=\exp\left(-\left\|x_i-x_j\right\|^2/(2\sigma^2)\right)$.

(4) Multilayer perceptron kernel function: $K(x_i,x_j)=\tanh\left(\gamma(x_i,x_j)+c\right)$, γ, c are constants, and the decision function of the final model is:

$$y(x)=\mathrm{sgn}\left[\sum_{i=1}^{N}\alpha_iy_iK(x_i,x_j)+b\right] \tag{22}$$

where b is the classification threshold.

2.3.2 LS-SVM

The conventional SVM framework demands considerable computational resources due to the necessity of solving a constrained quadratic optimization problem, where the number of constraints grows proportionally with the size of the training samples. Least Squares Support Vector Machine (LS-SVM) is an enhanced variant of the standard support vector machine model, in which LS-SVM replaces the inequality constraints in the SVM model with equality constraints and replaces the slack variable ξ_i with the square of the training error e_i^2, thus transforming the solution of the quadratic programming problem into finding a solution to a set of linear equations, the improved model is as follows:

$$\begin{aligned} &Minimize\ \frac{1}{2}\|w\|^2+\frac{C}{2}\sum e_i^2 \\ &S.t.\ y_i\left(w\cdot x_i+b\right)=1-e_i,i=1,2,\cdots,n \end{aligned} \tag{23}$$

The solution is also carried out by constructing a Lagrange function with the parameter to be solved in the decision function $\alpha=\left[\alpha_1,\cdots,\alpha_n\right]^T$, and b can be found by the following equation:

$$\begin{bmatrix}0 & \vec{1}\\ \vec{1}^T & \Omega+C^{-1}I\end{bmatrix}\begin{bmatrix}b\\ \alpha\end{bmatrix}=\begin{bmatrix}0\\ y\end{bmatrix} \tag{24}$$

where the vector $\vec{1}=[1,1,\cdots,1]$, $y=[y_1,y_2,\cdots,y_n]^T$, and Ω is a $n\times n$ symmetric matrix:

$$\Omega_{ij}=\varphi(x_i)^T\varphi(x_j)=K(x_i,x_j)\ i,j=1,2,\cdots,n \tag{25}$$

Through this formulation, the training of LS-SVM essentially reduces to resolving a linear system of equations, substantially lowering the associated computational complexity.

3 Empirical results and analysis

3.1 Data preprocessing analysis

In this subsection, after determining the source of the dataset for this research, data preprocessing techniques are utilized to ensure the usability of the research dataset, and the next step is to verify the methodological desirability of the above preprocessing in terms of feature screening, outlier screening, and missing value processing, whereas the process of balancing the dataset, dataset slicing, and data standardization is relatively simple, and the analytical process will not be given in detail.

3.1.1 Characterization

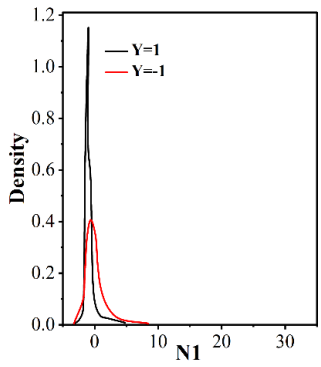
The Spearman correlation coefficient described in the preceding section is applied here to conduct tax risk management feature screening, and the feature screening results are shown in Table 1. Upon examining the tabulated correlation values, it is evident that the Spearman correlation coefficients of N10, N11, N12 are less than 0.8, there is no strong correlation, which is rejected, and the Spearman correlation coefficients of the other variables, namely N1, N2, N3, N4, N5, N6, N7, N8, N9 are all greater than 0.8, which satisfies the criteria established for the research standard, and finally after the tax risk management feature screening, nine feature variables are retained, which not only makes the feature variables more explanatory, but also improves the performance of the model. 9 feature variables are retained, which not only makes the feature variables more interpretable, but also improves the performance of the model.

Table 1: Feature screening results

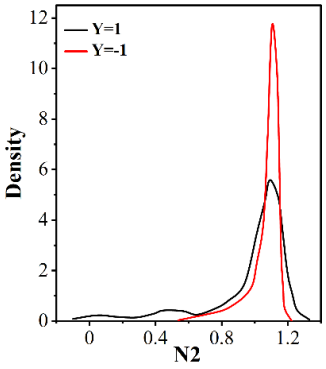
Name	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12
N1	1	0.886	0.841	0.856	0.899	0.869	0.805	0.899	0.867	0.707	0.715	0.738
N2	0.886	1	0.817	0.812	0.832	0.817	0.859	0.827	0.862	0.748	0.703	0.797
N3	0.841	0.817	1	0.803	0.897	0.805	0.882	0.833	0.864	0.722	0.765	0.718
N4	0.856	0.812	0.803	1	0.862	0.835	0.854	0.897	0.807	0.771	0.736	0.758
N5	0.899	0.832	0.897	0.862	1	0.813	0.822	0.817	0.822	0.797	0.761	0.713
N6	0.869	0.817	0.805	0.835	0.813	1	0.885	0.846	0.816	0.733	0.744	0.751
N7	0.805	0.859	0.882	0.854	0.822	0.885	1	0.834	0.881	0.701	0.723	0.702
N8	0.899	0.827	0.833	0.897	0.817	0.846	0.834	1	0.853	0.795	0.735	0.792
N9	0.867	0.862	0.864	0.807	0.822	0.816	0.881	0.853	1	0.781	0.756	0.746
N10	0.707	0.748	0.722	0.771	0.797	0.733	0.701	0.795	0.781	1	0.722	0.712
N11	0.715	0.703	0.765	0.736	0.761	0.744	0.723	0.735	0.756	0.722	1	0.751
N12	0.738	0.797	0.718	0.758	0.713	0.751	0.702	0.792	0.746	0.712	0.751	1

In order to more clearly grasp the distributional characteristics of the data, this paper pairs the use of kernel density plots to observe the distributional patterns of the data. For continuous

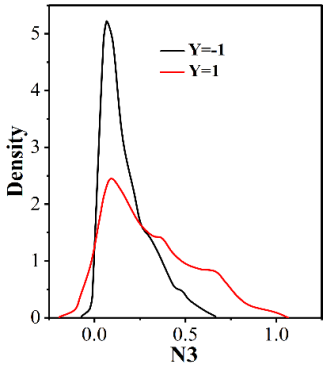
variables, kernel density plots are drawn on the basis of categorical grouping of the dependent variable Y to reflect the difference of each variable between the tax crisis sample (Y=1) and the tax normal sample (Y=-1), and the distribution of the characteristic data is shown in Fig. 1, in which (a)~(i) denote N1~N9, respectively.The horizontal coordinates of the plots represent the homogeneous segments of the indicator values, and the left vertical coordinates represent the probability density, which can be used to represent the probability distribution of a certain set of data and the probability of occurrence of a particular value, and the right vertical coordinate represents the number of observations (frequency) within each segment. Obviously, the above figure shows that N1 (cash ratio), N2 (ratio of tangible assets), N3 (ratio of fixed assets), N4 (ratio of current assets), N5 (ratio of non-current assets), N6 (turnover of current assets), N7 (basic earnings per share), N8 (diluted earnings per share), and N9 (net cash content of operating income) all appear to be the number of times of the data that the left and right sides of the distribution curve are asymmetric, and all of them have a long-tailed phenomenon, with an overall skewed distribution. Among them, N4 (current assets ratio) reflects the proportion of short-term assets relative to total asset holdings. In the financial normal sample, there are more firms between 0.7 and 0.8, which indicates that the firms' assets are highly liquid and have sufficient ability to repay short-term debts. Whereas, among the financial crisis firms, there are more current asset ratios between 0.304 and 0.517, which are prone to insolvency. This suggests that IT firms should try to increase their current asset ratios to 0.7 or above, and if they are less than 0.7, managers should pay attention to the risk of a tax crisis and adjust the composition of assets and liabilities.



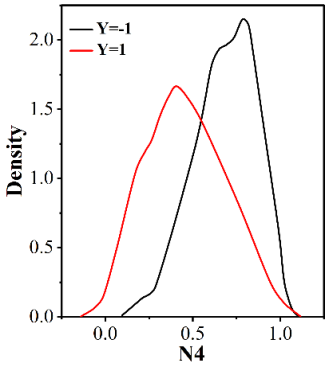
(a)N1



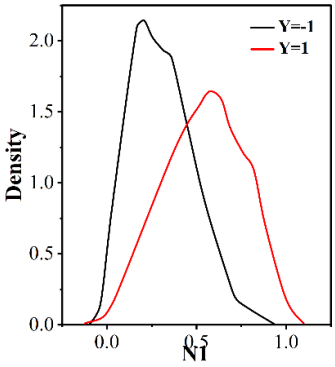
(b)N2



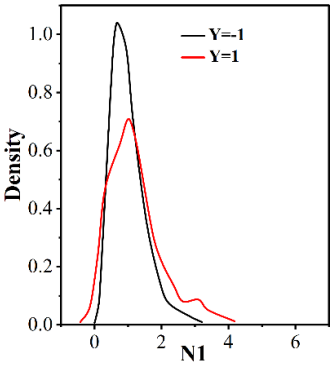
(c)N3



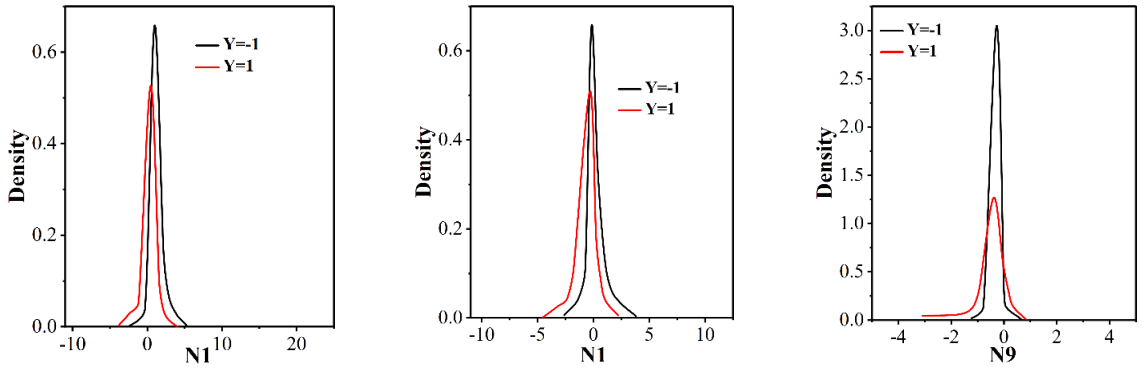
(d)N4



(e)N5



(f)N6

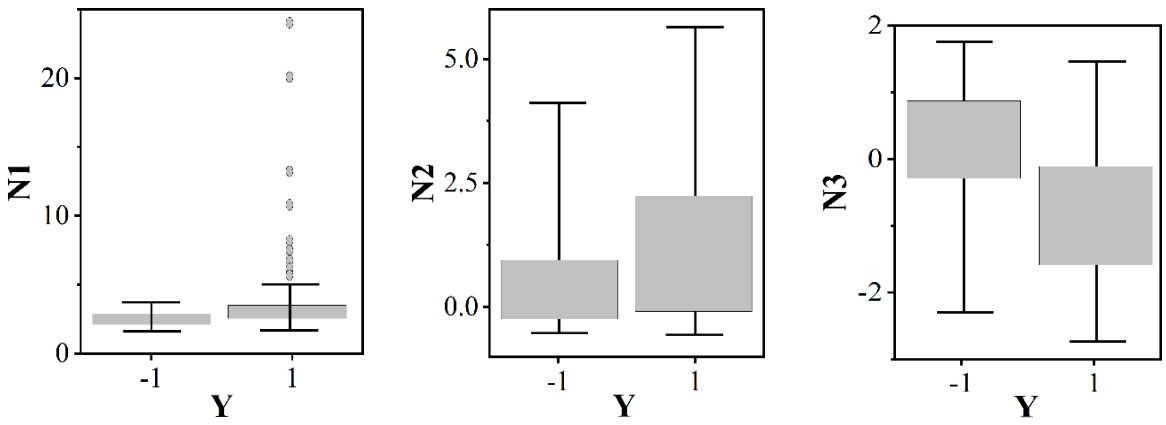


(g)N7 (h)N8 (i)N9

Figure 1: Distribution of characteristic data

3.1.2 Analysis of outlier handling

In data analysis, box plot identification is highly respected as a rapid and efficient method for filtering abnormal data. Its core lies in the use of interquartile range (IQR) to describe the distribution of data. Data samples can be considered to show a normal distribution within the interval bounded by the upper and lower thresholds, and data beyond this range will be defined as outliers. Among them, IQR denotes the spread between the first and third quartiles, i.e. the gap between the upper and lower quartiles, while c is an adjustable parameter. When c is set to 1.5, it is considered a mild outlier. When c is set to 3, it is considered an extreme outlier. Figure 2 shows the box-and-line plot of the sample data, where y=-1 represents the box-and-line plot of the financial normal sample, y=1 represents the box-and-line plot of the financial crisis sample, and the dots are the outliers. For the treatment of outliers, it is not possible to simply replace them, but to further analyze the reasons for the abnormal data. If it is a problem of data storage or other technical aspects, review and check first, and if it is an unreliable sample belonging to a tax fraud company, it is directly deleted. In the scenario of tax crisis early warning studied in this paper, the abnormal values of the financial indicators of the normal sample (y=-1) are filled with the median after deletion, and the financial indicators of the crisis sample (y=1) are in a special situation themselves, and their abnormal values are very valuable, which can effectively reflect the abnormalities of the company's tax status, and are conducive to the early warning of the tax crisis, then they can be retained.



(a)N1 (b)N2 (c)N3

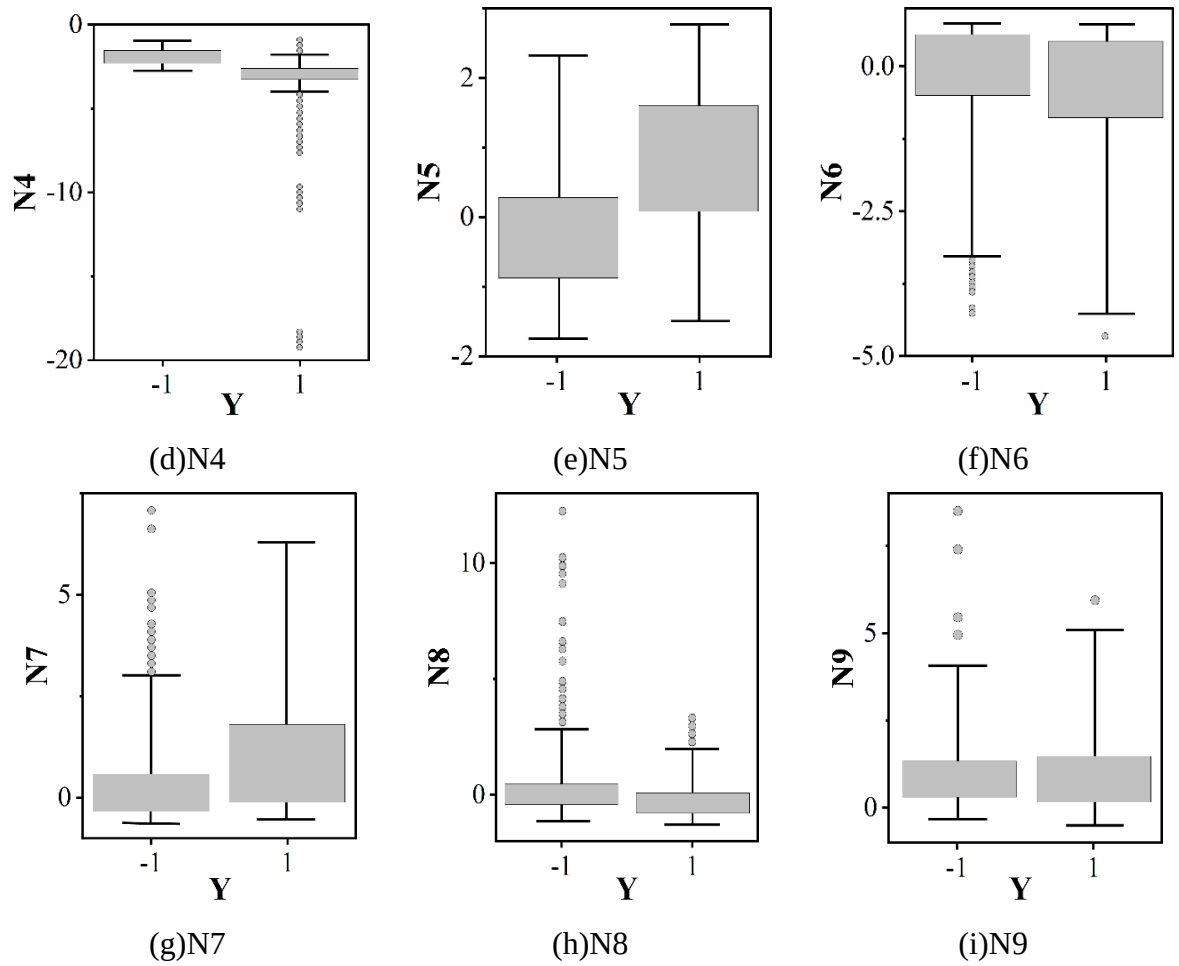


Figure 2: Box plot of the sample data

3.1.3 Missing value processing analysis

An interpolation-based strategy is adopted to impute the absent values across the above 9 tax risk features, so that its data availability can be further improved, and the processing analysis of the missing values of tax risk features is shown in Figure 3. After filling in the interpolation method, the missing rate of the nine tax risk features is controlled below 0.01, which ensures that the above selected features can accurately reflect the tax risk of listed enterprises.

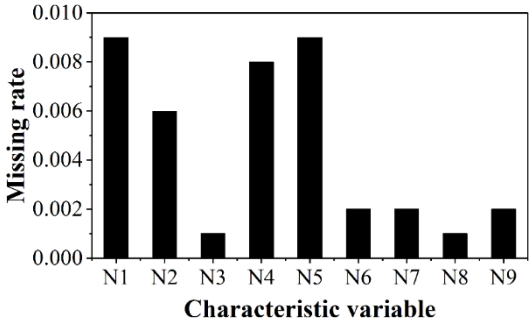


Figure 3: Handling of missing values in tax risk characteristics

3.2 Model validation analysis

Based on the preprocessing of the above dataset, the form of constructing confusion matrix is

utilized to validate the tax risk management identification model based on LS-SVM, with a view to providing theoretical guidance for the exploration of the application of artificial intelligence in tax risk identification and management in the context of big data.

3.2.1 Confusion matrix

In order to reflect the priority of LS-SVM hybrid algorithm, this paper selects decision tree algorithm, random forest algorithm, logistic regression algorithm as comparison algorithm from commonly used tax risk management identification.

The nine feature variables selected above are substituted into the decision tree algorithm, and the identification results of 0 and 1 indicate that the enterprise tax normal samples and enterprise tax risk samples, respectively, and the confusion matrix of the decision tree algorithm is shown in Table 2, and the ROC curve of the decision tree algorithm is shown in Figure 4. There are 2329 correct predictions for enterprise tax normal and 212 predictions for enterprise tax normal as enterprise tax risk, 804 correct predictions for enterprise tax risk and 208 predictions for enterprise tax risk as enterprise tax normal, the total number of predicted enterprise tax normal samples amounts to 2541, and the total count of predicted enterprise tax risk samples reaches 1021. Based on the accuracy rate, precision rate, recall rate, F1 value, and AUC value, the following metrics are derived:

(1) Correctness: $accuracy = \frac{TP + TN}{TP + FP + FN + TN} = 0.8818$.

(2) Precision rate: $\Pr ecision = \frac{TP}{TP + FP} = 0.9166$.

(3) Recall: $\mathrm{Re}\, call = \frac{TP}{TP + FN} = 0.918$.

(4) F1 value: $F1 = \frac{2 \times precision \times recall}{precision + recall} = 0.9173$.

(5) AUC value: AUC=0.9398.

Table 2: Confusion matrix of decision tree algorithm

Predicted value				
True value		0	1	Total
	0	2329	208	2537
	1	212	804	1016
	Total	2541	1021	3553

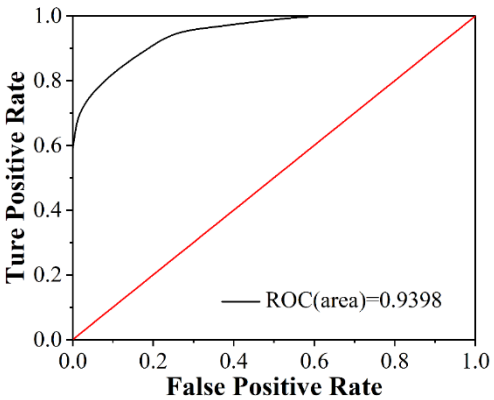


Figure 4: The ROC curve of the Quick Strategy Tree algorithm

The confusion matrix corresponding to the Random Forest algorithm is presented in Table 3, and its ROC curve is illustrated in Figure 5. There are 2431 correct predictions for corporate tax normal and 219 predictions for corporate tax normal as corporate tax risk, 777 correct predictions for corporate tax risk and 126 predictions for corporate tax risk as corporate tax normal, the aggregate number of samples classified as corporate tax normal is 2541, and the aggregate count of samples classified as corporate tax risk is 1021. According to the accuracy rate, precision rate, recall rate, F1 value, and AUC value, the following metrics are derived:

(1) Correctness: $accuracy = \frac{TP + TN}{TP + FP + FN + TN} = 0.9029$.

(2) Precision rate: $\Pr ecision = \frac{TP}{TP + FP} = 0.9174$.

(3) Recall: $\mathrm{Re}call = \frac{TP}{TP + FN} = 0.9507$.

(4) F1 value: $F1 = \frac{2 \times precision \times recall}{precision + recall} = 0.9337$.

(5) AUC value: auc=0.9398.

Table 3: The confusion matrix of the random forest algorithm

Predicted value				
True value		0	1	Total
	0	2431	126	2537
	1	219	777	1016
	Total	2541	1021	3553

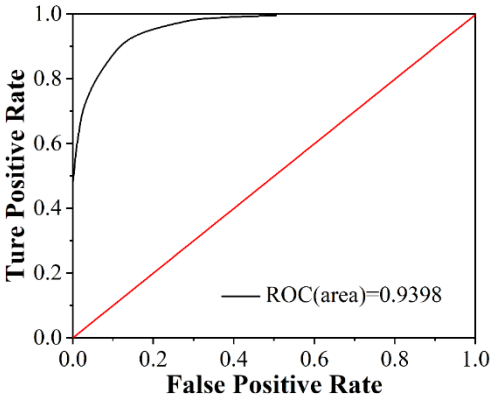


Figure 5: The ROC curve of the random forest algorithm

The confusion matrix for the logistic regression algorithm is displayed in Table 4, and the corresponding ROC curve appears in Figure 6. There are 2442 correct predictions for corporate tax normal and 227 predictions for corporate tax normal as corporate tax risk, 780 correct predictions for corporate tax risk and 104 predictions for corporate tax risk as corporate tax normal, the aggregate number of samples classified as corporate tax normal is 2669, and the aggregate count of samples predicted as corporate tax risk is 884. According to the accuracy rate, precision rate, recall rate, F1 value, and AUC value, the following metrics are derived:

(1) Correctness: $accuracy = \frac{TP + TN}{TP + FP + FN + TN} = 0.9068$.

(2) Precision rate: $\Pr ecision = \frac{TP}{TP + FP} = 0.9149$.

(3) Recall: $\mathrm{Re}\, call = \frac{TP}{TP + FN} = 0.9591.$

(4) F1 value: $F1 = \frac{2 \times precision \times recall}{precision + recall} = 0.9365$.

(5) AUC value: auc=0.9893.

Table 4: The confusion matrix of the logistic regression algorithm

Predicted value				
True value		0	1	Total
	0	2442	104	2546
	1	227	780	1007
	Total	2669	884	3553

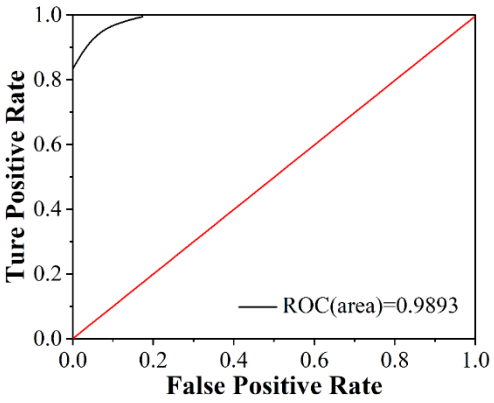


Figure 6: The ROC curve of the logistic regression algorithm

The confusion matrix of the LS-SVM algorithm is presented in Table 5, with its ROC curve depicted in Figure 7. There are 2556 correct predictions for corporate tax normal and 58 predictions for corporate tax normal as corporate tax risk, 916 correct predictions for corporate tax risk and 23 predictions for corporate tax risk as corporate tax normal, and the aggregate number of samples classified as corporate tax normal is 2614, and the aggregate count of samples predicted as corporate tax risk is 939. According to the formula of accuracy rate, precision rate, recall rate, F1 value, and AUC value the following metrics are derived:

(1) Correctness: $accuracy = \frac{TP + TN}{TP + FP + FN + TN} = 0.9772$.

(2) Precision rate: $\Pr ecision = \frac{TP}{TP + FP} = 0.9778$.

(3) Recall: $\mathrm{Re}\, call = \frac{TP}{TP + FN} = 0.9911.$

(4) F1 value: $F1 = \frac{2 \times precision \times recall}{precision + recall} = 0.9844$.

(5) AUC value: auc=0.9922.

Table 5: The confusion matrix of the logistic regression algorithm

Predicted value				
True value		0	1	Total
	0	2556	23	2579
	1	58	916	974
	Total	2614	939	3553

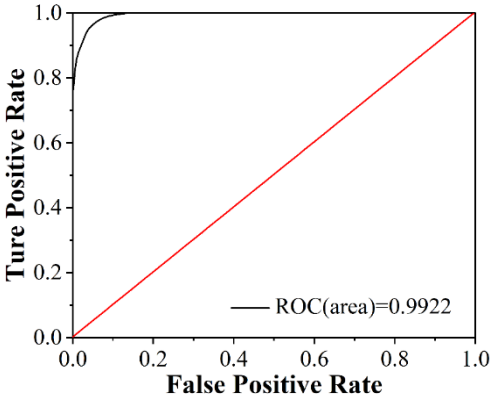


Figure 7: The ROC curve of the LS-SVM algorithm

3.2.2 Results and analysis

The above has been carried out by decision tree, random forest, logistic regression, LS-SVM for enterprise tax risk management identification, and the accuracy rate, precision rate, recall rate, F1 value, and AUC of each model have been computed via confusion matrix evaluation, and the ROC curve diagram has been generated. The accuracy rate, precision rate, recall rate, F1 value and AUC area of each model were summarized and compared and the histogram of each evaluation index was drawn, and the comparative visualization of performance metrics across all models is presented in Figure 8. Combined with the values in the figure, it can be seen that LS-SVM has great superiority in tax risk management identification, whether it is the accuracy rate, precision rate, recall rate, F1 value or AUC area, it consistently outperforms the other three models, and it is the optimal model among the four models. By comparison, the decision tree and random forest models demonstrate relatively inferior overall performance among the four models, while the logistic regression model, though outperforming the decision tree and random forest, is still a certain gap from the LS-SVM, which fully proves the value of the application of the LS-SVM algorithm in tax risk identification and management.

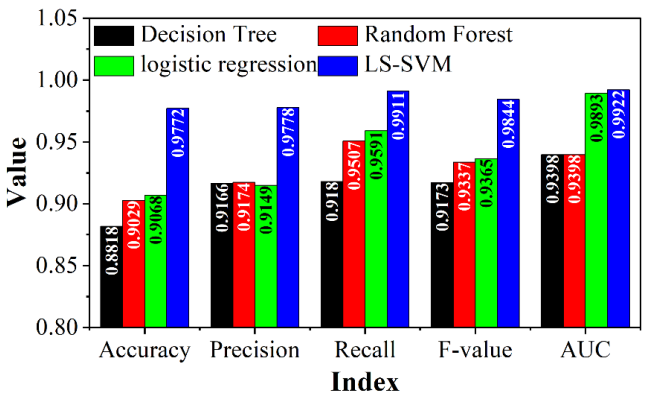


Figure 8: Comparison of evaluation indicators of each model

4 Conclusion

For the purpose of achieving intelligent tax risk identification and management, artificial intelligence technology in the context of big data undoubtedly provides broader opportunities for change, but also brings certain risk challenges. The source of the research dataset is first determined, and the interfering information in the dataset is removed through preprocessing, and then the hybrid LS-SVM algorithm is utilized to construct the enterprise tax risk management identification model. Finally, the empirical analysis of this paper's model is carried out by combining the dataset and evaluation indexes.

(1) Among the 12 enterprise tax risk management features in the original dataset, the Spearman correlation coefficients of N10, N11, and N12 are less than 0.8, and N10, N11, and N12 are excluded from the processing, while the remaining 9 items meet the research requirements, and are incorporated as independent input variables into the tax risk management identification model based on LS-SVM, so that the results of the analysis of the subsequent research will have stronger interpretability.

(2) The hybrid LS-SVM algorithm has great superiority in tax risk management identification, and its values are better than those of the decision tree algorithm, random forest algorithm, and logistic regression algorithm, which demonstrates the value of the application of the LS-SVM algorithm in tax risk identification and management.

About the Author

Xiaowen Shen was born in Heze, Shandong, China in 1989. She graduated from Yunnan Minzu University with a master's degree and is currently teaching at Sichuan Polytechnic University. Her main research directions are intelligent finance and taxation, and digital economy.

References

[1] Jin, Y., Li, Q., & Lu, B. (2023). Research on the Tax Risks of E-commerce Enterprises in the Big Data Environment. Journal of Humanities, Arts and Social Science, 7(6).

[2] Ouyang, S., & Fang, Y. (2022). Enterprise Financial and Tax Risk Management Methods under the Background of Big Data. Mathematical Problems in Engineering, 2022(1), 5831866.

[3] RUDENKO, V., & POHRISHCHUK, H. (2024). Management of tax risks of business entities. World of finance, (3 (80)), 121-134.

[4] Rathi, A., Sharma, S., Lodha, G., & Srivastava, M. (2021). A study on application of artificial intelligence and machine learning in indian taxation system. Psychology and Education Journal, 58(2), 1226-1233.

[5] Xavier, O. C., Pires, S. R., Marques, T. C., & Soares, A. D. S. (2022). Tax evasion identification using open data and artificial intelligence. Revista de Administração Pública, 56, 426-440.

[6] Nuryani, N., Mutiara, A. B., Wiryana, I. M., Purnamasari, D., & Putra, S. N. W. (2024). Artificial intelligence model for detecting tax evasion involving complex network schemes. Aptisi Transactions on Technopreneurship (ATT), 6(3), 339-356.

[7] Li, H. (2020). Modeling method of tax management system based on artificial intelligence. International Journal on Artificial Intelligence Tools, 29(07n08), 2040023.

[8] Belahouaoui, R., & Attak, E. H. (2024). Digital taxation, artificial intelligence and Tax Administration 3.0: improving tax compliance behavior–a systematic literature review using textometry (2016–2023). Accounting Research Journal, 37(2), 172-191.

[9] Dwianika, A., Sofia, I. P., & Retnaningtyas, I. (2023). Tax Compliance: development of artificial intelligence on tax issues. KnE Social Sciences, 728-733.

[10] Aina, A. T. (2024). Perceptions of Nigerian Tax Officers and Stakeholders on the Adoption of Artificial Intelligence in Tax Risk Management. FUDMA Journal of Accounting and Finance Research [FUJAFR], 2(4), 122-136.

[11] Hossain, M. Z., Hasan, L., Kumu, R. A., Bepari, M., & Sultana, S. (2025). The Role of Artificial Intelligence in Taxation and Compliance: Challenges and Future Prospects. EJSMT, 1(6), 73-85.

[12] Braun Binder, N. (2019). Artificial intelligence and taxation: risk management in fully automated taxation procedures. In Regulating Artificial Intelligence (pp. 295-306). Cham: Springer International Publishing.

[13] Rahman, S., Sirazy, M. R. M., Das, R., & Khan, R. S. (2024). An exploration of artificial intelligence techniques for optimizing tax compliance, fraud detection, and revenue collection in modern tax administrations. International Journal of Business Intelligence and Big Data Analytics, 7(3), 56-80.

[14] He, X. (2025). Research on Tax Risk Identification and Assessment System Assisted by Software and Deep Learning. International Journal of High Speed Electronics and Systems, 2540561.

[15] Huang, W., He, L., & Zhang, J. (2022, December). Artificial intelligence technology and tax risk management innovation. In International Conference on Computer, Artificial Intelligence, and Control Engineering (CAICE 2022) (Vol. 12288, pp. 362-366). SPIE.

[16] Bezditnyi, V. (2024). Use of artificial intelligence for tax planning optimization and regulatory compliance. Research Corridor Journal of Engineering Science, 1(1), 103-142.

[17] ALmusaway, M. A. K., Al-Tobi, B. H. M., & Kadhm, A. J. (2025). Examining the Role of Artificial Intelligence in Auditing for Tax Gap Reduction. Asian Journal of Economics, Business and Accounting, 25(5), 473-483.

[18] Saragih, A. H., Reyhani, Q., Setyowati, M. S., & Hendrawan, A. (2023). The potential of an artificial intelligence (AI) application for the tax administration system's modernization: the case of Indonesia. Artificial Intelligence and Law, 31(3), 491-514.

[19] Zhelev, Z. (2024). Application of AI to minimize information risk in tax control: Evidence from Bulgaria. Edelweiss Applied Science and Technology, 8(6), 5066-5074.

[20] Antón, F. S. (2021). Artificial intelligence and tax administration: strategy, applications and implications, with special reference to the tax inspection procedure. World Tax J.,

575.

[21] Raikov, A. L. E. X. A. N. D. E. R. (2021). Decreasing tax evasion by artificial intelligence. IFAC-PapersOnLine, 54(13), 172-177.

[22] Bajpai, D. A. (2024). Evaluating the impact of artificial intelligence on enhancing tax compliance and financial regulation. Available at SSRN 4922459.

[23] Rahayu, P. (2024). The impact of artificial intelligence on taxation aspect: A qualitative study. InFestasi, 20(1), 38-53.

[24] Han, N., Xu, W., Song, Q., Zhao, K., & Xu, Y. (2025). Application of interpretable artificial intelligence for sustainable tax management in the manufacturing industry. Sustainability, 17(3), 1121.